**ROJAS Luis - ECO372 Assignment 3**

**Rojassa1 – 1003650676**

***Exercise 1***

***Question a***.

The author states that the outcome of interest is college attendance and completed schooling, also the variable of interest is the eligibility for Social Security student benefits. Moreover, the author points out that to capture the effect of aid an exogenous source of variation is needed. In this case, changes in aid policies that affects some students while others remain unaffected.

This approach is more efficient that the approach that follows equation 1, since the latter does not capture the true casual effect of aid eligibility on school attendance rates. To capture the true casual effect the author uses the source of exogenous variation, this helps to realize the differences in students that are similar in an aspect that may affect the eligibility for a program and evaluates their outcomes when the program is available versus when it is not.

Meanwhile, the traditional approach fails to capture the real effect of aid, since aid eligibility is sometimes related to observed and unobserved characteristics that lead the decision of school. This may bias the outcome upwards or downwards, so this approach fails to reflect the true effect of aid on school decisions.

***Question b.***

```
. tab yearsr

    Year in |
    which a |
     senior |      Freq.     Percent        Cum.
------------+-----------------------------------
         79 |        933       23.41       23.41
         80 |      1,030       25.84       49.25
         81 |        919       23.06       72.30
         82 |        869       21.80       94.10
         83 |        235        5.90      100.00
------------+-----------------------------------
      Total |      3,986      100.00

.
. gen SrBeforeElimination = c.yearsr <= 81 //create a dummy equal 1 if graduated when benefits available

. label variable SrBeforeElimination "Graduated in Year where Benefits Available"

.
. tab SrBeforeElimination

  Graduated |
    in Year |
      where |
   Benefits |
  Available |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,104       27.70       27.70
          1 |      2,882       72.30      100.00
------------+-----------------------------------
      Total |      3,986      100.00
```

### *Question c.*

```
. table SrBeforeElimination [weight=wt88], by(fatherdec) contents (mean coll mean hgc23) //This creates a table where means are columns and
>  the 2 other variables rows
(frequency weights assumed)

----------------------------------------------
Father deceased by   |
age 18 and           |
Graduated in Year    |
where Benefits       |
Available            |   mean(coll)  mean(hgc23)
---------------------+------------------------
Father not deceased  |
                   0 |     .4756936     13.25053
                   1 |     .5017017     13.41341
---------------------+------------------------
Father deceased      |
                   0 |     .3522178     12.90348
                   1 |     .5604556     13.44166
----------------------------------------------
```

| | ***Mean of School Attendance*** | | ***Mean of Years of Schooling*** | |
|---|---|---|---|---|
| | ***Father Deceased*** | ***Father No Deceased*** | ***Father Deceased*** | ***Father No Deceased*** |
| ***Grad with Benefits*** | 0.56045 | 0.5017 | 13.44 | 13.41 |
| ***Grad without Benefits*** | 0.3522 | 0.4756 | 12.90 | 13.25 |

### *Question d.*

```
. reg coll offer##fatherdec [weight=wt88], cluster(hhid)
(analytic weights assumed)
(sum of wgt is 1,302,933,368)

Linear regression                               Number of obs     =      3,986
                                                F(3, 3122)        =       2.19
                                                Prob > F          =     0.0875
                                                R-squared         =     0.0020
                                                Root MSE          =     .49973

                             (Std. Err. adjusted for 3,123 clusters in hhid)
------------------------------------------------------------------------------
                |               Robust
           coll |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
        1.offer |   .0260081   .0212723     1.22   0.222    -.0157011    .0677173
                |
      fatherdec |
Father deceased |  -.1234757   .0834565    -1.48   0.139    -.2871109    .0401595
                |
offer#fatherdec |
1#Father deceased |   .1822297   .0958771     1.90   0.057    -.0057589    .3702183
                |
          _cons |   .4756935    .018872    25.21   0.000     .4386907    .5126964
------------------------------------------------------------------------------

. reg coll SrBeforeElimination##fatherdec [weight=wt88], cluster(hhid)
(analytic weights assumed)
(sum of wgt is 1,302,933,368)
```

| | (1)<br>Difference-in-Differences |
|---|---|
| Before=1 | 0.0260<br>(0.0213) |
| Father deceased | -0.123<br>(0.0835) |
| Before=1 # Father deceased | 0.182<br>(0.0959) |
| Constant | 0.476***<br>(0.0189) |
| Observations | 3986 |
| R-squared | 0.00196 |
| F-statistic | 2.187 |

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

### *Question e.*

The key assumption of this Differences in Differences is that children of deceased fathers will change their behaviour with respect to school attendance due to the removal of Student Benefits.

Also, this diff-in-diff only captures the intention to treat since the author focus on eligibility for the student benefits not only in the receipt of those.

The author runs a regression where she uses two set of interaction terms. The first one uses the covariates (family size, income, parental education, and marital status of household head, AFTQ scores, age, race, and gender) with the fact if the student was graduated in a cohort where she can access the Student Benefits. The second set uses the first set interacting with the deceased father dummy. As the author points out the interaction absorbs the bias caused by heterogeneity across time and eligibility status.

The author finds out that when the students with deceased fathers were using the student benefits, their rate of college attendance was almost the same, in magnitude, to the peer with living fathers (56.00 vs 50.2). However, when the student benefits were removed these rates fall by almost 20 percent, for those with deceased fathers, while their peers remained with almost the same rate (47.6). This estimate is statistically significant at the 6% level.

***Exercise 2***

***Question a.***

In this context $\kappa_1$ is the average hourly wage increase for a worker that is resident in a city where a big headquarters of a company is settle down in a determined year. In other words, a worker in a city with a big company's headquarters in a specific year, earns $\kappa_1$ more on average in comparison with workers not in this specific city.

***Question b.***

The coefficient $\kappa_2$ represents the average change in hourly wages in each year from 2012 to 2018 for workers working in city A. Meanwhile, the coefficient $\kappa_3$ represents the average change in hourly wages in each year from 2012 to 2018 for workers working in city B.

***Question c.***

```
. gen POST = c.year >= 2016 //Creates a dummy variable =1 for observations in year 2016+

. label variable POST "Dummy variable for observations in year 2016 and following"

. tab POST //to know the distribution and exactness of the new variable

      Dummy |
   variable |
        for |
observation |
  s in year |
   2016 and |
  following |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      4,000       57.14       57.14
          1 |      3,000       42.86      100.00
------------+-----------------------------------
      Total |      7,000      100.00
```

```
. gen HQ = cityB*POST //this interaction variable allow us to know the amount of observations in city B after the 2016, where the new headq
> uarter came to town

. label variable HQ "City with BIG Headquarters, 2016+"

. tab HQ

  City with |
        BIG |
Headquarter |
  s, 2016+ |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      5,800       82.86       82.86
          1 |      1,200       17.14      100.00
------------+-----------------------------------
      Total |      7,000      100.00
```

***Question d.***

```
.         //QUESTION D
. gen u = rnormal(0, 1.5) if cityA == 1
(2,800 missing values generated)

. label variable u "White Noise, different for each city"

. replace u = rnormal(0,1) if cityB == 1
(2,800 real changes made)
```

***Question e.***

```
. gen ys2012 = year-2012

. label variable ys2012 "years passed since 2012"

. tab ys2012

      years |
     passed |
 since 2012 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,000       14.29       14.29
          1 |      1,000       14.29       28.57
          2 |      1,000       14.29       42.86
          3 |      1,000       14.29       57.14
          4 |      1,000       14.29       71.43
          5 |      1,000       14.29       85.71
          6 |      1,000       14.29      100.00
------------+-----------------------------------
      Total |      7,000      100.00
```

***Question f.***

```
.
.       //QUESTION F
. gen w = 10 +1.3 * (HQ) + 0.2 * (ys2012*cityA) + 0.6 * (ys2012*cityB) + u // generating variable w with the given coefficients

. label variable w "Hourly Wages with the given coefficients"

.
.
```

***Question g.***

```
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
workerID        float   %9.0g
cityA           byte    %8.0g                 Number of workers in city A
cityB           byte    %8.0g                 Number of workers in city B
year            float   %9.0g
POST            float   %9.0g                 Dummy variable for observations in year 2016 and following
HQ              float   %9.0g                 City with BIG Headquarters, 2016+
u               float   %9.0g                 White Noise, different for each city
ys2012          float   %9.0g                 years passed since 2012
w               float   %9.0g                 Hourly Wages with the given coefficients
-------------------------------------------------------------------------------
```

Drop the variables for Big Headquarters and the Error term

```
. drop (HQ u )

. describe

Contains data
  obs:         7,000
 vars:             7
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
workerID        float   %9.0g
cityA           byte    %8.0g                 Number of workers in city A
cityB           byte    %8.0g                 Number of workers in city B
year            float   %9.0g
POST            float   %9.0g                 Dummy variable for observations in year 2016 and following
ys2012          float   %9.0g                 years passed since 2012
w               float   %9.0g                 Hourly Wages with the given coefficients
-------------------------------------------------------------------------------
```

***Question h.***

$$w_{it} = \kappa_0 + \kappa_1 POST + \kappa_2(cityB) + \kappa_3(POST * cityB)$$

***Question i.***

The hourly wage of workers who work in a city where exist Big Headquarters of a firm, like Facezom, will be 1.3 higher on average, this happens after 2016.

***Question j.***

```
.
. reg w cityB##POST, robust

Linear regression                               Number of obs     =      7,000
                                                F(3, 6996)        =    3450.66
                                                Prob > F          =     0.0000
                                                R-squared         =     0.5165
                                                Root MSE          =     1.3641

------------------------------------------------------------------------------
             |               Robust
           w |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.cityB |    .566042    .042403    13.35   0.000     .4829192    .6491648
      1.POST |   .7432085   .0466444    15.93   0.000     .6517714    .8346456
             |
  cityB#POST |
        1 1  |   2.730324   .0634366    43.04   0.000     2.605969    2.854679
             |
       _cons |   10.27439    .030471   337.19   0.000     10.21466    10.33412
------------------------------------------------------------------------------
```

In this case the coefficient, is inflated and too different to the coefficient in the expected regression for this case.

This coefficient overestimates the effect of work in a city with big headquarters on hourly wages. This coefficient is only based on the observations after 2016 and the fact that Facezon is settling down on city B. However, we know that 60% of the observations come from city A, and only 40% come from city B.

For this reason, we need to analyze a broad set of dates since we know we have 7000 observations. In other words, this is a case of Selection Bias, so analyzing the data after 2016 will only focus on 42.86% of the observations and this may drive up the coefficient in POST. If we do not control for the whole set of observations our error term will be giant.

***Question k.***

```
. reg w POST##cityB c.ys2012##cityB, robust

Linear regression                               Number of obs     =      7,000
                                                F(5, 6994)        =    2683.14
                                                Prob > F          =     0.0000
                                                R-squared         =     0.5581
                                                Root MSE          =     1.3043

------------------------------------------------------------------------------
             |               Robust
           w |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      1.POST |   .0299584   .0912417     0.33   0.743     -.148903    .2088198
     1.cityB |   .0003292   .0589231     0.01   0.996    -.1151779    .1158363
             |
  POST#cityB |
         1 1 |   1.410327   .1186689    11.88   0.000       1.1777    1.642954
             |
      ys2012 |   .2037857   .0225522     9.04   0.000     .1595765     .247995
             |
cityB#c.ys2012 |
           1 |   .3771419   .0294297    12.82   0.000     .3194508    .4348329
             |
       _cons |    9.96871   .0454355   219.40   0.000     9.879643    10.05778
------------------------------------------------------------------------------
```

$$w_{it} = \kappa_0 + \kappa_1 POST + \kappa_2 cityB + \kappa_3(POST * cityB) + \kappa_4(ys2012) + \kappa_5(ys2012 * cityB)$$

The best way to remedy this regression will be to know how the hourly wages in each city evolves over time. Since we have the variables required, we can make another interaction term (ys2012##cityB) to know how these hourly wages changed over time from city to city. This in comparison with the old regression from question h, will absorb better the changes in wages and decrease the error term. Then, we can have a better perspective of the casual effect of Big Headquarters on an hourly wage.

According to this regression, the hourly wage of workers who work in a city where there is a Big Headquarter of a firm like Facezom after 2016, will be 1.41 higher on average.

In comparison with *question i,* the average earning for people living in such cities is 0.11 higher. This is due to the dropping of the ***u*** variable, where the u variable is the error term. According to the data generation process, the error term is different for each city, but both errors are random numbers from a normal distribution of mean 0 for both. The main difference is in the standard deviation for each city, where cityA has a standard deviation of 1.5 and cityB has a standard deviation of 1.

Moreover, the t-test for the hypothesis testing $H_a: \beta_1 = \beta_2$ versus $H_1: \beta_1 \neq \beta_2$ reveals that the t-value is equal to $\frac{1.41-1.3}{0.1186689} = 0.927$, so we fail to reject the null hypothesis even at the 10% level and conclude that the coefficients are equal. This means that this regression is statistically significant and in majority captures the true casual effect of Big Headquarters on the hourly wages of workers.